Deep Learning Fusion of RGB and Depth Images for Pedestrian Detection

Zhixin Guo
zhixin.guo@ugent.be
Wenzhi Liao
wenzhi.liao@ugent.be
Yifan Xiao
xiao.yifan@ugent.be
Peter Veelaert
peter.veelaert@ugent.be
Wilfried Philips
wilfried.philips@ugent.be

Department of Telecommunications and Information Processing
Ghent University-IMEC
Gent, Belgium

Abstract

In this paper, we propose an effective method based on the Faster-RCNN structure to combine RGB and depth images for pedestrian detection. During the training stage, we generate a semantic segmentation map from the depth image and use it to refine the convolutional features extracted from the RGB images. In addition, we acquire more accurate region proposals by exploring the perspective projection with the help of depth information. Experimental results demonstrate that our proposed method achieves the state-of-the-art RGBD pedestrian detection performance on KITTI [12] dataset.

1 Introduction

Pedestrian detection is a key task in many vision based applications, such as surveillance and autonomous driving. Despite the fact that significant progress has been made in recent years [3, 7, 8, 36], there is still room for improvement [38], especially under challenging conditions such as partial occlusions and insufficient illumination. Such conditions, however, are difficult to handle with RGB images alone.

Recent work has shown that depth images could be used as a good complementary to RGB images [16, 19, 25, 26]. Unlike RGB images, depth images will not provide fine appearance details of the objects, but rather present more clear silhouettes of them. In addition, depth images show a better description of the relationship of the spatial positions of the objects in a scene, which enriches the representations of each target object. For example in Fig. 1, if we apply a cluster analysis on the depth region of the pedestrian in the green bounding box, we can clearly distinguish the pedestrian from the occluding object in front of it and the background behind it. For depth images generated from other sensors (e.g. lidar), they are also not affected by bad lighting conditions, which seriously harms the quality of RGB images from visible cameras.

Figure 1: An RGB and depth image pair from KITTI dataset. With a cluster analysis of the depth region in the bounding box, the person is distinguished from the occluding object and the background, which can be utilized to boost the detection performance.

Looking back the development of RGBD pedestrian detection researches, depth information was used as additional hand-crafted features in traditional computer vision techniques. Detectors trained with the combination of depth and RGB features showed an improved performance [11, 33, 34]. After the rise of Convolutional Neural Networks (CNNs) [18, 23, 29], most RGBD object detection frameworks apply two parallel CNN network streams for RGB and depth modalities, and fuse them in early or late layers [9, 10, 16, 19, 25]. However, in those work, RGB and depth images are processed in the same way, which neglects the different characteristics of RGB and depth image modalities. Some researchers take advantage of the depth images to reduce the searching space of candidate regions [2, 21], but the spatial relationship illustrated in Fig. 1 is still not well exploited.

In this paper, we focus on how to better utilize the depth images for pedestrian detection. We demonstrate that by refining RGB features with the guidance of depth information, the detection performance can be significantly improved, even without directly extracting features from depth images. As far as we know, this is different from all the existing RGBD pedestrian detection frameworks, which equally feed the network with RGB and depth images and train them together. Our major contribution is threefold:

- First, we propose to use depth image to guide the reweighting of the convolutional features extracted from RGB images. It helps the classification network to pay more attention on the features from the target pedestrian instead of on the non-pedestrian regions (e.g. occluding objects, background). This framework shows an significant improvements on pedestrian detection, both for the visible pedestrians and partially occluded ones.

- Second, by exploring the projection relationship with depth information, we generate better candidate regions during the region proposal stage, which further improves the framework of our RGBD detector.

- Third, our proposed RGBD detector significantly outperforms the baseline method Faster R-CNN and the state-of-the-art RGBD detection algorithms on KITTI dataset.

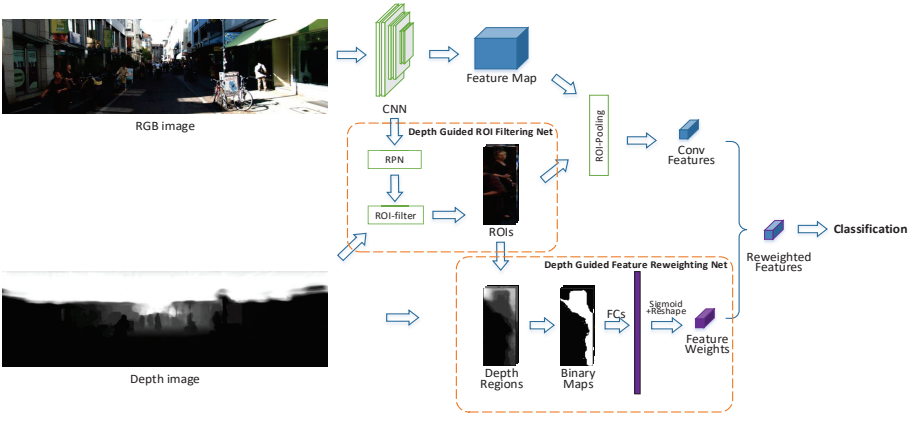
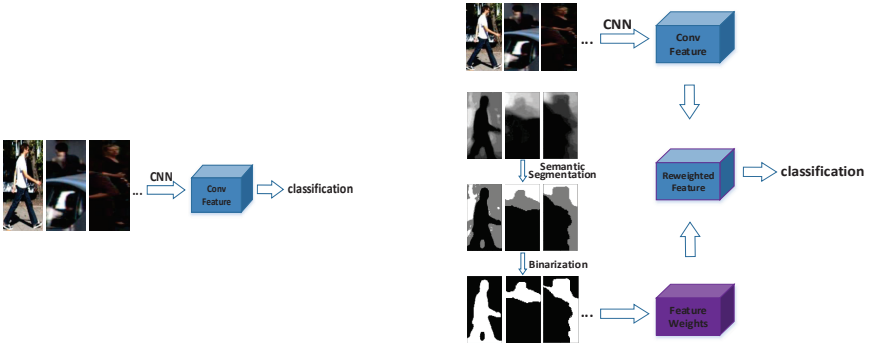


Figure 2: The architecture of our proposed RGBD pedestrian detector. It is based on a Faster-RCNN structure with two subsets: (1) Depth Guided Feature Reweighting Net, which produces the weights for the convolutional feature of each ROI. (2) Depth Guided ROI Filtering Net, which evaluate each ROI with the perspective projection relationship. The reweighted feature are used in classification.

2 Related Work

CNN based pedestrian detection: With the rise of CNNs, significant improvements have been made in object detection tasks [13, 14, 18, 23, 27, 29]. One pioneer work used deep CNNs for pedestrian detection was proposed in [28], which combined CNNs with unsupervised multi-stage feature learning. Some works [1, 20, 31, 32] follow the RCNN [14] structure, which integrates external proposals with CNN framework and outperforms the traditional hand-crafted feature based pedestrian detectors [8, 24] by a significant margin. Later, Fast-RCNN [13] and Faster-RCNN [27] further improves the RCNN framework by sharing computations during feature extraction and region proposal generation. Some adaptations of the Faster-RCNN framework has reached state-of-the-art results in pedestrian detection [36, 37]. In this paper, an adapted Faster-RCNN framework which takes RGB and depth images as inputs is used in our experiments.

Pedestrian detection in RGBD images: Early works [6, 21, 30] extended hand-crafted features from RGB data with depth information to provide a richer representation. [30] and [15] designed HOG-like features in depth channel and trained a SVM classifier for pedestrian detection. Depth is also used to reduce the search space of candidate regions [9, 21, 35]. In most CNN based methods, features are extracted from RGB and depth images independently and fused together for further classification. Gupta *et al.* [16] proposed to encode depth image into a three-channel (height above ground, horizontal disparity and angle to gravity) depth image, and used a CNN network pre-trained on RGB images to extract depth features. Tanguy *et al.* [25] exhaustively trained several models to explore the optimal fusion layer of RGB and depth features in CNN networks. Gupta *et al.* [17] transfered supervisions from RGB images to depth images to achieve a richer representation. Instead of equally extracting features from RGB and depth images, our method takes advantage of the depth modality to refine the RGB features, which outperforms the existing RGBD detection methods.



(a) typical CNN classification process (b) CNN classification with reweighted features

Figure 3: The comparison of the typical CNN classification process with our method. We reweight the convolutional features according to the depth information, which increase the contribution of features from the pedestrian.

3 Framework

An overview of our framework is illustrated in Fig. 2. It is based on the Faster-RCNN structure [27]: A CNN network (e.g. VGG16 net) takes an RGB image as input and extracts the convolutional feature of the whole image. Then a region proposal network (RPN) is used to generate regions of interest (ROIs). After cropping the features of each ROI from the feature map, a ROI-Pooling layer is applied to generate same-length features for each ROI. Then the features will go through a classification network, usually some fully connected layers, to produce the classification results and bounding box regression informations.

Two subnets are added to the Faster-RCNN structure to take advantage of the depth image: Depth Guided Feature Reweighting Net and ROI Filtering net. The depth region of each ROI is put into the reweighting net, where a binary map is produced to indicates the pixels from the target pedestrian and guides the reweighting of the RGB convolutional features. Depth images are also used in the ROI generation process: In the depth guided ROI filtering net, the size of each candidate region will be evaluated according to the perspective projection relationship with the help of the depth image. ROIs which have reasonable size will be kept for classification, while others are removed.

3.1 Depth Guided Feature Reweighting

Typically, machine learning based detectors treat all the pedestrians equally. For example in Fig. 3(a), pedestrian samples with variant appearances (e.g. different postures, occlusion conditions, backgrounds...) will go through a CNN network to generate a n-dimensional feature vector F and produce a classification score, which can be computed in Eq. 1:

$$Score = \sum \omega_i F_i, i = 0, 1, ..., n \tag{1}$$

where ω_i represents the weight of the classification network (here we equivalent the classification network as a 1-layer fully connected network), which determines the contribution of feature F_i to the classification score. For different pedestrian samples, F_i could either come from the pedestrian region, or come from non-pedestrian regions such as the background

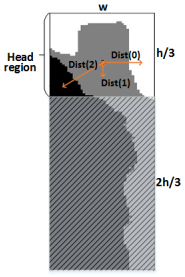


Figure 4: Computing the average distance between the head center and the pixels from layer m. Only pixels from the head region are taken into account.

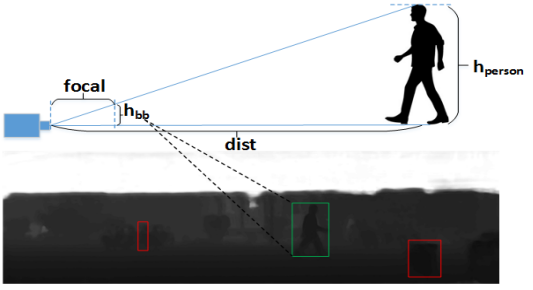


Figure 5: The projection relationship of a camera. We can compute the distance between the pedestrian and the camera through depth information, and further estimate the height of the pedestrian.

or the occluding object. Therefore, the network should be adaptive to different pedestrian samples, by increasing ω_i to enhance the contribution of F_i which comes from the pedestrian region, while decreasing ω_i to reduce the influence of F_i which comes from the non-pedestrian region. However, ω_i is fixed after training, so we propose to reweight F_i instead. The flowchart of the reweighting net is illustrated in Fig. 3(b). The depth image is used to guide the generation of feature weights in three steps:

Semantic Segmentation: We apply a 3-means clustering for the depth region of each ROI to segment it into three parts. Normally the top depth layer, which is nearest to the camera and has the smallest depth value, represents the pedestrian (see the first sample in Fig. 3(b)). But when the pedestrian is partially occluded (see the second and third samples in Fig. 3(b)), the occluding object becomes the top depth layer, while the pedestrian and background occupy middle layer and bottom layer, respectively.

Binarization: The goal of binarization is to generate a binary map which marks the pixels from the pedestrian region. To realize this, we propose to choose the depth layer which is nearest to the center of the head region, which we define as the top 1/3 of each ROI, see Fig. 4. Note that we only consider the pixels from the head region, which has the smallest probability of being occluded. $(\frac{w}{2}, \frac{h}{6})$ represents the center point of the head region, where w and h indicates the width and height of the ROI region, respectively. $Dist(m)$ is defined to represent the average euclidean distance between the head center and all the pixels from each depth layer m, which is computed in Eq. 2:

$$Dist(m) = \frac{1}{N^m} \sum_{i=1}^{N^m} \sqrt{(x_i^m - \frac{w}{2})^2 + (y_i^m - \frac{h}{6})^2}, m = 1,2,3 \tag{2}$$

where (x_i^m, y_i^m) indicates the location of the ith pixel from the mth depth layer, while N^m indicates the total pixel numbers from layer m. Pixels from the layer with the smallest $Dist(m)$ will be set as 1, while other pixels are all set to 0.

Weights Learning: The binary maps will go through two fully connected layers to generate the weight vector, which is followed by a sigmoid computation and a reshape operation to get the final weight cube. It will have the same size with the convolutional feature. After a element-wise multiplication, the reweighted feature will be sent to the classification network.

Note that, similar to our scheme, Zhang *et al.* [39] proposed an attention net to guide the reweighting of convolutional features. However, it only focuses on the occlusion condi-

tions, which uses part detectors to generate heat maps and reweight the feature by channel to increase the contribution of visible body parts. In our method, we take advantage of the depth map to better distinguish the pedestrian with not only the occluding object, but also the background, which improves the detection performances of both fully visible and partially occluded pedestrians. In addition, we also use the depth map to eliminate unreasonable candidates during the generation of ROIs, which is explained in Sec. 3.2.

3.2 Depth Guided ROI Filtering

In typical Faster-RCNN structure, an RPN is used to generate ROIs for further classification. In order to make sure true positives will not be lost in this stage (or they will not be recovered any more), the number of ROIs is normally large. On one hand, it harms the efficiency of the detector. On the other hand, some high score false positive ROIs are still difficult to be recognized during the classification. In this section we propose to take advantage of the depth information to efficiently screen some unreasonable ROIs.

As is shown in Fig. 5, the height of the bounding box h_{bb} can be computed by Eq. 3:

$$h_{bb} = \frac{focal}{dist} \cdot h_{person} \tag{3}$$

where $dist$ represents the distance between the pedestrian and camera, while h_{person} represents the real height of the pedestrian. According to the depth value, $dist$ can be computed by comparing with the maximum range of depth map:

$$dist = \frac{depth}{255} \cdot range \tag{4}$$

where the $depth$ value of the pedestrian is computed by the same method described in Section 3.1. In Eq. 3, we assume the range of h_{person} to be [1m, 2m], which should include most of the pedestrian heights in real life. The bounding boxes with a height out of this range are considered to be unreasonable, and removed from the ROI candidates.

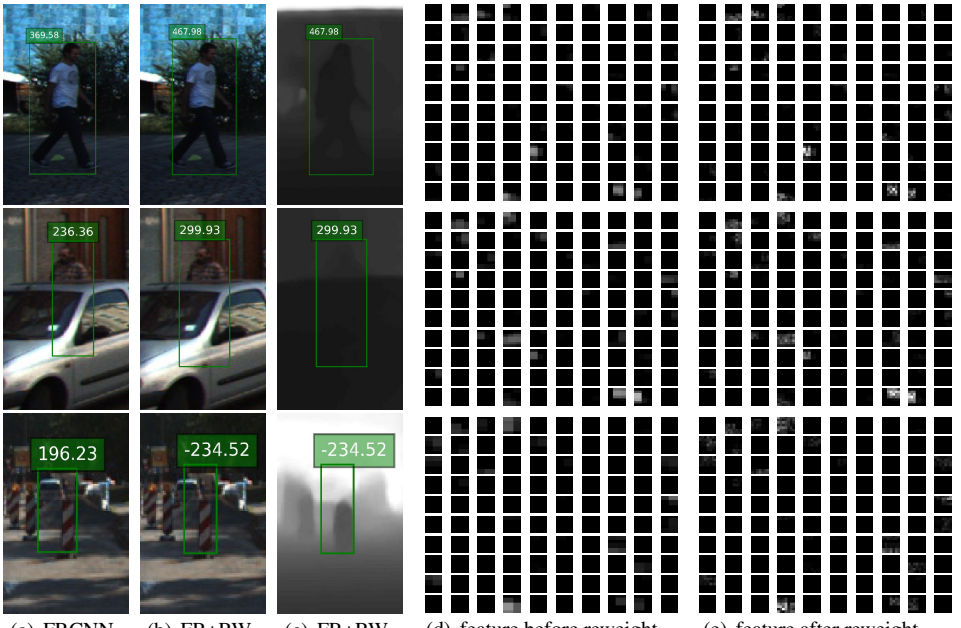
4 Experiments

In this section, we will first introduce the implementation details. Then we will show some experiments of the two subnets in our framework. In the end, we will compare our method with the baseline method and some state-of-the-art RGBD detection methods.

4.1 Implementation Details

Our RGBD pedestrian detection framework is based on an adapted version of Faster-RCNN: In order to better fit for pedestrians, we use 4 aspect ratios of [1, 1,5, 2, 2.5] and 5 scales of [1, 2, 4, 8, 16] in the RPN, which allows more human-like regions and small regions during the ROI generation. We use the VGG16 net pre-trained on ImageNet as the feature extraction network, but the last max-pooling layer is removed, which reduced the feature stride from 16 pixels to 8 pixels to further improve the detection of small pedestrians. To evaluate the performance of our adapted Faster-RCNN detector, we apply a 3-fold cross validation in the KITTI training set, which shows the average precision (AP) improvements from 61.5% to 67.3%. This adapted Faster-RCNN will be one of our baseline method.

The depth images are generated from [6], which is computed from the lidar cloud points.



(a) FRCNN (b) FR+RW (c) FR+RW (d) feature before reweight (e) feature after reweight

Figure 6: Three detection samples from FRCNN and our FR+RW method. After the feature reweighting guided by the depth image, our FR+RW method generates more distinguishable classification scores for true positive and false positive detections.

4.2 Experiments on Feature Reweighting

In order to evaluate our feature reweighting method, we compare the detection results of our baseline Faster-RCNN (denoted as FRCNN) structure and the Faster-RCNN with our Depth Guided Feature Reweighting Net (denoted as FR+RW).

Fig. 6 shows three detection samples, including a fully visible pedestrian, a partially occluded pedestrian and a false detection. We can find that after a depth-guided feature reweighting, our FR+RW method significantly increases the classification score of the true positive detections (row 1 and row 2), while reduces the classification scores of the false positive detection (row 3). We attribute it to larger contributions of the features from the pedestrian region, while the feature reweighting does not help the non-pedestrian detections in a reasonable way. The convolutional features before and after reweighting are visualized in Fig. 6(d) and Fig. 6(e). Here we show the feature channels 162-262 from the 512 channel features extracted from the VGG16 net. More detection results of multiple pedestrians are shown in Sec. 4.4.

4.3 Experiments on ROI Filtering

In default settings of FRCNN, the RPN will select at most 300 ROIs for further classification, which requires much computation. We compare the top 30 ROIs generated from default RPN and RPN with our ROI filtering network (denoted as FRCNN+filter) in Fig. 7. We can find that with the ROI filtering network, the generated ROIs will be mostly focused on the pedestrian regions. It allows us to use fewer ROIs during the region proposal generation. In

(a) FRCNN (b) FRCNN+filter

Figure 7: The top 30 ROIs generated from the RPN of FRCNN and our FRCNN+filter method. The ROIs generated from our method has higher quality, which are mostly located at pedestrian regions.

our following experiments, we choose to use the top 100 ROIs for classification.

4.4 Comparison with other methods

Here we compare our proposed method with several baseline methods:

FRCNN: Our adapted version of Faster-RCNN (see Sec. 4.1 for detailed settings). This method is used as our baseline RGB pedestrian detector and denoted as FRCNN.

FRCNN+dep: In order to make a fair comparison with our method which also uses depth images, we train two parallel network streams to extract features from both RGB and depth modalities and fuse them in later layers. We compared different fusion points for all the five convolutional layers in VGG16 net and found fusion after layer-3 performed best, which was consistent with the results in [22, 25]. So in our experiment, we fuse the layer-3 RGB and depth features by adding a concatenation layer, which follows with a 1x1 convolution filter to connect to the fourth convolutional layer. This structure is used as our baseline RGBD pedestrian detector and denoted as FRCNN+dep.

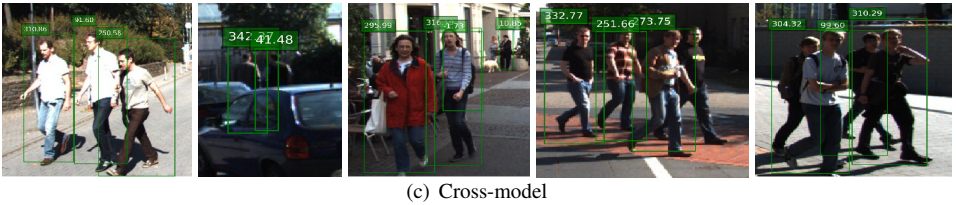
Cross-modal: In our experiments we also compare with the state-of-the-art RGBD object detection method [17]. The middle layer features from RGB modality is used as supervision for learning rich representations from depth modality. It is used as another baseline RGBD detector and denoted as Cross-modal.

Proposed: Our RGBD pedestrian detector illustrated in Fig. 2, which is based on the FRCNN structure and contains two subnets: Depth Guided Feature Reweighting Net and ROI Filtering Net.

We evaluate the above methods on KITTI dataset. The detection results are evaluated based on three levels of difficulty (Easy, Moderate and Hard), which are defined by different height, occlusion level and truncation percentage of the pedestrians. Since the ground truth labels of the test set are not available, in our experiments we randomly split the training set (7481 images) into three sets and apply a 3-fold cross-validation for evaluation.

(a) FRCNN

(b) FRCNN+dep



(c) Cross-model

(d) Proposed

Figure 8: The comparison results of FRCNN, FRCNN+dep, Cross-model and our proposed method. The proposed method outperforms the other methods, especially under occlusion conditions.

Table 1 shows that by naively fusing RGB and depth features in CNN networks, the FRCNN+dep method does not achieve an obvious improvement. We attribute this to the characteristics of the KITTI dataset, because most of the images are captured in daytime and with sufficient luminance, which diminishes the advantage of the depth modality. Then we find the state-of-the-art RGBD detection method Cross-modal achieves a significant improvement in all the three test cases, while our proposed method further outperforms Cross-modal by 0.9%, 3.2% and 8.1% in Easy, Moderate and Hard test cases, respectively. It shows that our method works well under more difficult conditions (e.g. occlusion), which is also proved in the detection results in Fig. 8.

5 Conclusion

In this paper, we propose an RGBD pedestrian detection framework, which takes advantage of depth images to guide the refining of convolutional features extracted from RGB images. In addition, the depth image is also used to guide the filtering of unreasonable ROIs according to the projection relationships. We report state-of-the-art performance on KITTI dataset, which outperforms the existing RGBD object detection methods, especially under occlusion

Method	Easy	Moderate	Hard
FRCNN	72.8%	67.3%	61.8%
FRCNN+dep	74.2%	66.8%	63.7%
Cross-modal	82.6%	72.6%	65.1%
Proposed	**83.5%**	**75.8%**	**73.2%**

Table 1: Comparison between the four methods. Our proposed method outperforms the others, especially in Moderate and Hard cases.

conditions.

Acknowledgements: This work is supported by the Chinese Scholarship Council (CSC), and the grant number is 201606220043.

References

[1] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. Real-time pedestrian detection with deep network cascades. 2015.

[2] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool. Pedestrian detection at 100 frames per second. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2903–2910. IEEE, 2012.

[3] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision*, pages 613–627. Springer, 2014.

[4] Massimo Camplani, Adeline Paiement, Majid Mirmehdi, Dima Damen, Sion Hannuna, Tilo Burghardt, and Lili Tao. Multiple human tracking in rgb-depth data: a survey. *IET computer vision*, 11(4):265–285, 2016.

[5] Wongun Choi, Caroline Pantofaru, and Silvio Savarese. Detecting and tracking people using an rgb-d camera via multiple detector fusion. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 1076–1083. IEEE, 2011.

[6] Martin Dimitrievski, Peter Veelaert, and Wilfried Philips. Learning morphological operators for depth completion. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 450–461. Springer, 2018.

[7] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012.

[8] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (8):1532–1545, 2014.

[9] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015.

[10] Christian Ertler, H Posseger, M Optiz, and Horst Bischof. Pedestrian detection in rgb-d images from an elevated viewpoint. In *22nd Computer Vision Winter Workshop*, 2017.

[11] Dariu M Gavrila and Stefan Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International journal of computer vision*, 73(1):41–59, 2007.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[15] Domingo Iván Rodríguez González and Jean-Bernard Hayet. Fast human detection in rgb-d images with progressive svm-classification. In *Pacific-Rim Symposium on Image and Video Technology*, pages 337–348. Springer, 2013.

[16] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360. Springer, 2014.

[17] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016.

[20] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4073–4082, 2015.

[21] Omid Hosseini Jafari, Dennis Mitzel, and Bastian Leibe. Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5636–5643. IEEE, 2014.

[22] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 49–56, 2017.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[24] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2014.

[25] Tanguy Ophoff, Kristof Van Beeck, and Toon Goedemé. Exploring rgb+ depth fusion for real-time object detection. *Sensors*, 19(4):866, 2019.

[26] Cristiano Premebida, Joao Carreira, Jorge Batista, and Urbano Nunes. Pedestrian detection combining rgb and dense lidar data. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4112–4117. IEEE, 2014.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[28] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann Lecun. Pedestrian detection with unsupervised multi-stage feature learning. In *Computer Vision and Pattern Recognition*, 2013.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] Luciano Spinello and Kai O Arras. People detection in rgb-d data. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3838–3843. IEEE, 2011.

[31] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1904–1912, 2015.

[32] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5087, 2015.

[33] Ningbo Wang, Xiaojin Gong, and Jilin Liu. A new depth descriptor for pedestrian detection in rgb-d images. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3688–3691. IEEE, 2012.

[34] Shengyin Wu, Shiqi Yu, and Wensheng Chen. An attempt to pedestrian detection in depth images. In *2011 Third Chinese Conference on Intelligent Visual Surveillance*, pages 97–100. IEEE, 2011.

[35] Hao Zhang, Christopher Reardon, and Lynne E Parker. Real-time multiple human perception with color-depth cameras on a mobile robot. *IEEE Transactions on Cybernetics*, 43(5):1429–1441, 2013.

[36] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European conference on computer vision*, pages 443–457. Springer, 2016.

[37] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017.

[38] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Towards reaching human performance in pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):973–986, 2018.

[39] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6995–7003, 2018.